Referential scales and case alignment: reviewing the typological evidence

Bickel, Balthasar ; Witzlack-Makarevich, Alena

Abstract: It has often been claimed that the distribution of case marking is systematically affected by a universal scale of referential categories. This can be understood as a universal correlation between the odds of overt case marking and scale ranks (a negative correlation for subjects, a positive one for objects), or as an implicational universal proposing that, if a language has a split in case marking, this split fits a universal scale. We tested both claims with various versions of scale definitions against a sample of over 350 case systems worldwide, controlling for confounding factors of genealogical and areal relationships. We find no statistical evidence for a universal correlation that is independent of family membership and has any appreciable predictive power. Formulated as an implicational universal, we find that there are only few areally independent families that show a trend towards fitting scales, and that each family fits different scales. What we do find, by contr ast, is a strong area effect: once genealogical relationships are controlled for, differential argument marking shows a frequency peak in Eurasia and nowhere else. We conclude that the currently available empirical evidence is too weak to reject the null hypothesis that splits in case marking develop through individual diachronic changes – such as innovations of case morphology in nouns but not pronouns (Filimonova, 2005), reanalyses of instrumentals as ergatives on inanimates (Garrett, 1990), contact-induced calquing of definite vs. indefinite contrasts by means of case marking, or other idiosyncracies.

Referential scales and case alignment: reviewing the typological evidence

Balthasar Bickel & Alena Witzlack-Makarevich*
University of Leipzig

Abstract

It has often been claimed that the distribution of case marking is systematically affected by a universal scale of referential categories. This can be understood as a universal correlation between the odds of overt case marking and scale ranks (a negative correlation for subjects, a positive one for objects), or as an implicational universal proposing that, if a language has a split in case marking, this split fits a universal scale. We tested both claims with various versions of scale definitions against a sample of over 350 case systems worldwide, controlling for confounding factors of genealogical and areal relationships. We find no statistical evidence for a universal correlation that is independent of family membership and has any appreciable predictive power. Formulated as an implicational universal, we find that there are only few areally independent families that show a trend towards fitting scales, and that each family fits different scales. What we do find, by contr ast, is a strong area effect: once genealogical relationships are controlled for, differential argument marking shows a frequency peak in Eurasia and nowhere else. We conclude that the currently available empirical evidence is too weak to reject the null hypothesis that splits in case marking develop through individual diachronic changes – such as innovations of case morphology in nouns but not pronouns (Filimonova, 2005), reanalyses of instrumentals as ergatives on inanimates (Garrett, 1990), contact-induced calquing of definite vs. indefinite contrasts by means of case marking, or other idiosyncracies.

*This research was supported by Grant No. BI 799/3-1 from the *Deutsche Forschungsgemeinschaft*. Bickel did the statistical analyses, theoretical interpretation and most of the write-up; Witzlack-Makarevich did most of the data analysis, database conceptualization and development. We are grateful to Taras Zakharko for programming a substantial portion of the data extraction and aggregation scripts that we used in the analyses and for many helpful suggestions. All computations were done in *R* (R Development Core Team, 2008), with the added packages 'Design' (Harrell, 2001), 'vcd' (Meyer et al., 2006), and 'MASS' (Venables and Ripley, 2002). We also thank Hans-Jörg Bibiko for an improvement of R's mapping functions. We would like to thank participants of the Workshop on theoretical morphology 3 for their comments and suggestions. We would also like to thank Tarald Taraldsen for his valuable comments.

Scales, 1-37
Marc Richards & Andrej L. Malchukov (eds.)
LINGUISTISCHE ARBEITS BERICHTE 86, Universität Leipzig 2008

1. Introduction

Typological generalizations are often first based on small-scale surveys or contrastive analyses of a few languages, and it is typically only later, after much additional empirical groundwork, that they can be evaluated through rigorous quantitative analysis. Many initial generalizations have been corroborated in this way over time (as is the case, for example, with most of Greenberg's word order correlations; Dryer, 1992), but other initial generalizations have turned out to be spurious (as is the case, for example, with claims about a principled distinction between 'agglutinating' vs. 'fusional' morphologies; Haspelmath, in press). Some initial generalizations, however, have never been subject to systematic and large-scale quantitative analysis. One such generalization is the idea that, universally, some kind of referential scale governs the kinds of case or adposition markings we find, such that, for example, first and second person pronoun stand a higher chan ce for accusative as opposed to ergative case marking.[1]

The idea was developed in the late 70s (Silverstein, 1976; Moravcsik, 1978; Comrie, 1981; DeLancey, 1981, among others) and despite the lack of large-scale empirical tests, it is now widely taken to be an established finding. Aissen (1999), for example, counts the idea "among the most robust generalizations in syntactic markedness" and accepts a version of the idea as reflecting an inviolable component of "universal grammar" (also cf. Kiparsky, 2004).

In this paper we subject the idea of scale effects on case marking to empirical testing against data from a large typological database with world-wide coverage. In order to do so, we first discuss various versions of the idea and reformulate them as precise and testable hypotheses (Section 2). In Sections 3 and 4, we subject these hypotheses to statistical tests, concluding in Section 5 that the empirical support for all hypotheses is surprisingly weak – much weaker indeed than for many other typological generalizations.

---

[1] In the following we use the term 'case' as a cover term for dependent-marking of argument roles, including adpositional marking and generalizing across the kind of morphology and phonology involved.

2. Claims and hypotheses

The idea of scale effects on case alignment does not easily translate into precise and testable hypotheses because there are many ways in which the idea can be spelled out – specifically, the hypotheses can be understood as absolute universals ('laws of grammar') or as probabilistic trends ('statistical universals'); as affecting overt case exponence (Comrie, 1981) or as affecting alignment in any kind of grammatical relation (Silverstein, 1976); as predicting the type of entire alignment or marking systems or as predicting correlations of alignment or marking systems with ranks on the scale. In the following we discuss these different ways of spelling out the basic idea.

2.1. Universals, variation, and exceptions

When hypothesized universals are shown to have exceptions, there are always two possible responses: one can try and 'explain away' the exceptions and thereby reduce the variation (i.e. choose a 'reductionist' approach); the hypothesized universal is then 'absolute', inviolable. Alternatively, one can measure the variation and try to explain it (i.e. choose a 'variationist' approach); the universal is then 'statistical' and violable to a degree that can be measured.

An example for a 'reductionist' approach is Kiparsky's (2004) tentative analysis of Arrernte: in Arrernte (e.g. Mparntwe Arrernte: Wilkins, 1989), first person singular pronouns and nouns have ergative case marking, all other pronouns show accusative alignment. Under a reductionist analysis, this unexpected distribution can be accounted for by claiming that despite their appearance, first person pronouns are nouns in this language, i.e. that they belong to the same part of speech as lexical nouns, while other pronouns constitute a part of speech of their own. The challenge for such an approach is of course to find independent evidence for the analysis. So far, we are not aware of any such evidence although we cannot obviously exclude the possibility of finding evidence. The intrinsic risk of the reductionist approach is non-testability because there is always a non-zero chance of discovering further apparent counterexamples of the Arrernte kind, and for these we cannot anticipate whether they can be explained away.

Under a 'variationist' approach, the Arrernte distribution counts as a real exception, and the question then is how many such exceptions there are, and whether they are less frequent than distributions that match the expectations. In this paper, we follow this variationist approach exclusively. The basic hy-

pothesis then is that there are universal principles of referential scale effects that 'push' the development of case distributions in certain ways. As a result, case distributions that fit the principles are predicted to be more common than others. The null hypothesis against which this prediction can be statistically tested, is that case distributions are not affected by universal principles of referential scale effects, but instead follow from what looks like random diachronic fluctuation, i.e. current case distributions follow from whatever diachronies they went through. For example, if an ergative arose from an instrumental, we expect it to be limited to inanimates. This will then mimic a referential scale effect, but under the null hypothesis, it will be a mere epiphenomenon cf. Garrett, 1990. Indeed, under the null hypothesis, it will just be as likely that, for example, an ergative case system decays in lexical nouns but survives in pronouns (cf. Filimonova, 2005). This will then lead to systems that do not mimic any referential scale effect and instead look like violations of such effects.

2.2. Marking, markedness, and alignment

Ever since its original formulations, the idea of scale effects has had two possible interpretations: under one interpretation (associated with Comrie, 1981), referential scales affect the distribution of overt case exponence: low-ranking A arguments and high-ranking O arguments are predicted to carry overt case markers ('ergative' and 'accusative', respectively) while high-ranking A and low-ranking O arguments are predicted to carry no overt case markers.[2] This can be extended to predictions on the phonological amount or morphological specification of case exponence, as in Keine and Müller's (2008) proposal in this volume.

An alternative interpretation (associated with Silverstein, 1976), makes predictions not about overt marking patterns but about abstract markedness relations: under this interpretation, low-ranking A arguments and high-ranking O arguments are predicted to be mapped into marked grammatical relations, while high-ranking A and low-ranking O arguments are predicted to be mapped into unmarked grammatical relations. The terms 'marked' and 'unmarked' are used in a classical structuralist sense in this approach and describe which grammatical relation is structurally more constrained or spec-

[2]We use A and O as symbols for proto-agent and proto-patient arguments in the sense of Dowty (1991). S stands for the sole argument of intransitives.

ified than the other. There are many technical ways in which the relevant constraints and specifications can be spelled out, but the one that is most often associated with Silverstein's original proposal has to do with the alignment of grammatical relations, i.e. the way arguments are mapped into sets (Bickel, in press-a). Given this, the relevant specifications are defined by alignment sets : the sets {S,O}, {S,A} and {S,A,O} are all less specific than the sets {A} and {O}. Therefore, we expect low-ranking A arguments and high-ranking O arguments to be associated with {A} and {O} relations, respectively, while high-ranking A and low-ranking O arguments are expected to be associated with sets that include S, (i.e. {S,A,O} or {S,A} for high-ranking A arguments, and {S,A,O} or {S,O} for low-ranking O arguments).

Silverstein's interpretation makes predictions for any kind of alignment set, i.e. any kind of grammatical relation. This includes not only alignment sets defined by case marking but also alignment sets defined by agreement systems, conjunction reduction, or whatever syntactic structures select specific arguments to the exclusion of others. Comrie's interpretation, by contrast, is limited to case marking. Bickel (in press-b) demonstrates that the generalization beyond case marking has no empirical support: tested against a world-wide database on alignment splits in agreement systems, there is no trend for such systems to follow the predictions. For alignments in other syntactic structures, we lack sufficiently rich databases, but a preliminary survey reveals no systematic trend either. For diatheses in particular, Bickel and Gaenszle (2007) show that there is no systematic association of scale ranks with passivization as opposed to antipassivization: first person O argument s, for example, are required to be passivized in just as many languages as they are required to be antipassivized. For grammatical relations targeted by relative clause constructions, there are both languages where higher-ranking arguments are preferred and languages where lower-ranking arguments are preferred (Bickel, in press-a).

With regard to case systems, Silverstein's and Comrie's versions make the same predictions to the extent that structurally unmarked relations tend to have less morphological exponence than structurally marked relations. We know of only one single language that deviates from this in having a morphologically marked {S,A} case, and shows at the same time an alignment split based on a referential scale: this is Middle Atlas Berber where the marked nominative is restricted to low-ranking S and A arguments. This fits Comrie's prediction that low-ranking A arguments receive morphologically overt marking. In return, it violates Silverstein's version of scale effects because low-ranking O arguments are mapped into a structurally marked grammatical

relation: O is mapped into the {O} set, which is structurally marked relative to the less specific {S,A} set. However, this is one language and we cannot make any statistical inferences from this.

Since there is no evidence for scale effects beyond case-marking and since for all but one relevant language, structural markedness correlates with morphological markedness, we focus on case (and adposition) marking and use structural markedness, i.e. alignment sets, as a proxy for morphological markedness.[3]

The only problematic case for this approach is presented by double-oblique alignment {A,O} vs. {S} that contrasts with ergative or accusative alignment. An example is Vafsi, a Northwestern Iranian language. In past tense clauses of this language, A arguments are in what is called the oblique case; O arguments are also in the same oblique case if they rank high in discourse status, e.g. by being definite (1a). Lower-ranking O arguments, by contrast, are in the 'direct' case (1b), which also covers S arguments (1c):

(1) Vafsi (Northwestern Iranian; Indo-European; Stilo, 2004)

a. luás-i kærg-é=s b̲æværdæ.
fox-OBL chicken-OBL.F=3s PUNCT-took
A O
'The fox took the chicken.'

b. in luti-an yey xær=esan æ-rúttæ.
DEM wise.guys-OBL.PL one donkey.DIR=3p DUR-sold
A O
'These wise guys were selling a donkey.'

c. zeni-́e há-nešesd-end.
woman-PL.DIR PVB-sat-3p
S
'The women sat down.'

Such a system sets up a contrast between {A,O} for high-ranking O arguments and {S,O} for low-ranking O arguments. Since the two alignments

[3]We do not choose the opposite route (using morphological exponence as a proxy for markedness) because determining the markedness of morphological exponence requires substantial additional research in morphophonology, which goes beyond our current project scope. Also, we suggest that any progress here will have to look into degrees of overt exponence, along the lines suggested by Keine and Müller (2008) in the present volume.

contain the same number of specifications (two each), one could argue that they are equally marked. However, closer inspection of the morphological markedness and what we know from the history of these languages (Haig, 2008) suggests that {A,O} represents the structurally marked forms. And since structural unmarkedness implies a more extensive distributional potential, the unmarkedness of an alignment set can just as well be defined in terms of whether or not the set contains an argument outside transitive verbs, i.e. S. In the following we assume this and define markedness directly in terms of alignment with S:

(2) An alignment set α is marked relative to another alignment set β iff α contains less argument roles than β and β contains S.

In the Vafsi example, this means that high-ranking O arguments are mapped into a marked alignment set (the {A,O} set), while low-ranking O-arguments are mapped into an unmarked set (the {S,O} set), in line with Silverstein's predictions.

Under these assumptions, hypotheses of scale effects are specifically about marked vs. unmarked argument sets: we expect marked sets to associate preferentially with low-ranking A and high-ranking O arguments. If there is no difference in markedness, then all ranks on the scale show the same distribution, and there is no prediction. This is the case in the Vafsi example with regard to the A argument: all A arguments are mapped into a marked alignment set, either {A} or {A,O}, and therefore always surface in the oblique case.

As the Vafsi data suggest, the predictions occasionally differ for A and O arguments, a difference enshrined in the traditional distinction between 'differential subject marking' and 'differential object marking'. Since all A arguments are marked, there is no prediction for A marking in Vafsi; for O arguments, by contrast, Vafsi is in line with the prediction that higher-ranking O have a higher chance of being marked than lower-ranking O arguments. While in this case there is a contrast between 'no prediction' and 'expected', some systems of alignment sets lead to conflicts in expectations. Khufi, an other Iranian language, restricts the double-oblique system to a subset of pronouns (first and second person singular, third person) and contrasts this with neutral alignment in all other NPs. The following data illustrate this: demonstrative (third person) pronouns are in the oblique case in A (3a) and O (3b) but not in S (3c) function; lex ical nouns are always in the direct case (cf. the O arguments in 3a and 3d, the A argument in 3d and the S argument in 3e):

(3) Khufi (Southeastern Iranian; Indo-European; Sokolova, 1959)

a. way xūðm wīnt.
DIST.SG.OBL dream.DIR see.PST
A O
'He saw a dream.'

b. mā́š=am way na talœpt.
1PL.DIR=1PL.PST DIST.SG.OBL NEG look.for.PST
A O
'We did not look for him.'

c. yaw yat tar dum yīd.
DIST.SG.DIR come.PST to MID.SG.OBL bridge.DIR
S
'He came towards that bridge.'

d. Tarsakbő́y žœr zů̊xt.
Tarsakboy.DIR stone.DIR take.PST
A O
'Tarsakboy took the stone.'

e. Tarsakbő́y ǰő́y-ti x̌āb na x̌ů̊vd.
Tarsakboy.DIR REFL place=on night NEG sleep.PST
S
'Tarsakboy did not sleep at his place that night.'

Such a distribution is expected for O arguments: only high-ranking (pronominal) O arguments are mapped into the marked {A,O} set; low-ranking O arguments are mapped into the unmarked {S,A,O} set. But for A arguments, the distribution is unexpected because high-ranking A arguments are also mapped into the marked set {A,O} set while low-ranking arguments are mapped into the unmarked {S,A,O} set.

There are many possibilities of how markedness sets distribute across referential scales. Table 1 illustrates some of these by data we have in our database. In Table 1 we simply divided the scale into 'high', 'mid' and 'low', and spell out the concrete scales out in the last column. But this begs the question of how referential scales are actually defined. We take this up in the following.

high	mid	low	prediction for A	prediction for O	example	relevant scale (segment) in example
{S,A}:{O}	{S,A,O}		none	many	Spanish	anim > inanim
{S,A}:{O}	{S,O}:{A}		many	many	Dyirbal (Dixon, 1972)	1/2 > 3/N
{S,A}:{O}	{S}:{A}:{O}	{S,O}:{A}	many	many	Djapu (Morphy, 1983)	Pro > N-hum > nonhum
{S,A,O}	{S,O}:{A}		many	none	Belhare (Bickel, 2003)	1s > 1d/1p/2/3/N
{S,A,O}	{S,A}:{O}		none	rare	Middle Atlas Berber (Pencheon, 1973)	1/2/3 > N
{S,O}:{A}	{S}:{A}:{O}	{S,O}:{A}	none	rare	Gumbaynggir (Eades, 1979)	3 > N-kin > N-other
{A,O}:{S}	{S,A,O}		rare	many	Khufi (Sokolova, 1959)	1s/2s/3 > 1p/2p/N
{A,O}:{S}	{S,A,O}	{A,O}:{S}	rare	rare	Vafsi past tense (Stilo, 2004)	1p/1s > 2p > 2s/3p/N
{A,O}:{S}	{S,O}:{A}		none	many	Vafsi past tense (Stilo, 2004)	N-high > N-low
{S}:{A}:{O}	{S,A}:{O}	{S}:{A}:{O}	rare	none	Talysh (Northern) past tense (Schulze, 2000)	1s > 2p/2s/3p > 3s
{S}:{A}:{O}	{S,O}:{A}		none	many	Nepali set I tense forms	anim/def > inanim/indef

Table 1: A selection of observed distributions of case alignment sets across referential scales ('none' means 'no prediction', 'many' means 'predicted to be frequent', 'rare' means 'predicted to be rare or non-existent')

2.3. Defining referential scales

A referential scale is a scale defined by referential categories, covering 'inherent' referential categories like 'animate', discourse-based referential categories like 'speaker' or 'proximative' and part of speech notions like 'pronoun'. Obviously, all these categories are ultimately language-specific and can only be identified by language-specific criteria (cf. Haspelmath 2008 in this volume). Yet, for many such categories, we can generalize over language-specific scales, because they show sufficient semantic overlap across languages. For example, it seems plausible that a category like 'first person singular' in one language is the same as the category 'first person singular' in another language. With categories like 'proximative' or 'topical', this is much less clear.

What is needed then is a list of category types that abstracts away from language-specific details and allows comparing language-specific referential categories, i.e. what is variously called 'typological types' (Bickel and Nichols, 2002), 'values of typological features' (Haspelmath et al., 2005), or 'comparative notions' (Haspelmath, 2007). Notions like 'proximative', 'topical', 'definite' etc., for example, are probably best captured by a typological type like 'higher discourse rank' which is defined in opposition to 'lower discourse rank', with the understanding the 'discourse rank' is a probabilistic notion determined by a series of factors whose weights may differ from language to language.

Such type lists can be declared *a priori*, or they can be derived inductively by generalizing over all and only those language-specific categories that are encountered. Most lists that have been proposed in the literature are probably developed on the basis of a mix of *a priori* expectations and experience gained through typological survey work. Generally recognized types include notions like first, second, and third person; singular vs. dual vs. plural; pronoun vs. lexical noun; definite/topical vs. indefinite/nontopical; human vs. (nonhuman) animate vs. inanimate (e.g. Comrie, 1981; Dixon, 1994; Croft, 1990). In our own database work we develop lists using the 'autotypologizing' method of Bickel and Nichols (2002): this method seeks to inductively abstract away from language-specific categories to exactly that degree that is needed to capture all language-specific distinctions encountered in a sample of language. After surveying 333 languages with this method, we find the list of types in Table 2 to be at the right level of abstraction for capturing all distinctions ever made by at least one language.

Given this list, the question is how it maps into a scale. It has often been

Type	definition
1duPro	1st person dual pronoun
1plPro	1st person plural pronoun
1sgPro	first person singular pronoun
2duPro	second person dual pronoun
2plPro	second person plural pronoun
2sgPro	second person singular pronoun
3duPro	third person dual pronoun
3plPro	third person plural pronoun
3plPro-high	third person pronouns plural with a higher discourse rank than '3plPro-low' (where rank is determined by discourse factors with language-specific weights)
3plPro-low	third person pronouns plural with la lower discourse rank than '3plPro-high' (where rank is determined by discourse factors with language-specific weights)
3sgPro	third person singular pronouns
3sgPro-high	third person pronouns singular with a higher discourse rank than '3sgPro-low' (where rank is determined by discourse factors with language-specific weights)
3sgPro-low	third person pronouns plural with a lower discourse rank than '3sgPro-high' (where rank is determined by discourse factors with language-specific weights)
3sg_humPro	third person singular pronoun with human reference
3sg_non-hum-Pro	third person singular pronoun with non-human reference
DEM	Demonstratives
N	lexical nouns, of any kind
N-anim	animate nouns
N-def	definite nouns
N-high	nouns with a higher discourse rank than 'N-low' (where rank is determined by discourse factors with language-specific weights)
N-high_anim	nouns denoting higher animates (humans and some animals)
N-hum	human nouns
N-inanim	inanimate nouns
N-indef	indefinite nouns
N-kin	kin terms
N-low	nouns with a lower discourse rank than 'N-high' (where rank is determined by discourse factors with language-specific weights)
N-low_anim	lower animates
N-non-hum-sg	non-human nouns in singular
N-non-kin	any noun apart from kin terms
N-non-sg	nouns in non-singular (i.e. N-pl and N-dual)
N-non-specific	nouns without specific reference
N-pl	Nouns in plural
N-sg	Nouns in singular
N-spec	nouns having specific reference
NOT-ProperN	all nouns apart from proper names
N_non-hum	non-human nouns
Nnon-pers	non personal nouns
PersN	personal names
Pro	free pronouns that head NPs (excluding pronominal agreement markers), of any kind
Pro-kin	kinship pronouns
ProperN	Proper name

Table 2: List of types needed to distinguish all categories relevant for case alignment sets in the 333 languages surveyed

noted that the details of scales vary from language to language – e.g. some languages rank first person above second person while others rank second person above first person – but that there still are some basic principles – e.g. that all languages rank speech act participants above third persons. There are many proposals in the literature on what exactly these basic principles

are, and in the following we explore an entire series of possible principles. In addition we also compute a best-fitting scale and explore this as well.

Label	definition
1>2>3>N	1plPro / 1sgPro / 1duPro > 2sgPro / 2plPro / 2duPro > 3sgPro / DEM / 3plPro / Pro-kin / 3duPro / 3sg_humPro / 3sg_non-hum-Pro / 3sgPro-high / 3plPro-high / 3sgPro-low / 3plPro-low > N / N-hum / ProperN / N-anim / N-kin / N-def / N-indef / N_non-hum / N-high_anim / N-low_anim / N-sg / N-pl / N-spec / N-non-specific / N-inanim / N-non-kin / N_specif&anim / PersN / N-non-sg / N-non-hum-sg / N-high / N-low / Nnon-pers / NOT-ProperN
SAP>3>N-high>N-low	2sgPro / 1plPro / 1sgPro / 2plPro / 1duPro / 2duPro > 3sgPro / DEM / 3plPro / Pro-kin / 3duPro / 3sg_humPro / 3sg_non-hum-Pro / 3sgPro-high / 3plPro-high / 3sgPro-low / 3plPro-low > N-hum / ProperN / N-anim / N-kin / N-def / N-high_anim / N-spec / PersN / N-high > N-indef / N_non-hum / N-low_anim / N-non-specific / N-inanim / N-non-kin / N-non-hum-sg / N-low / Nnon-pers / NOT-ProperN
SAP>3>N	2sgPro / 1plPro / 1sgPro / 2plPro / 1duPro / 2duPro > 3sgPro / DEM / 3plPro / Pro-kin / 3duPro / 3sg_humPro / 3sg_non-hum-Pro / 3sgPro-high / 3plPro-high / 3sgPro-low / 3plPro-low > N / N-hum / ProperN / N-anim / N-kin / N-def / N-indef / N_non-hum / N-high_anim / N-low_anim / N-sg / N-pl / N-spec / N-non-specific / N-inanim / N-non-kin / PersN / N-non-sg / N-non-hum-sg / N-high / N-low / Nnon-pers / NOT-ProperN
SAP>3/N	2sgPro / 1plPro / 1sgPro / 2plPro / 1duPro / 2duPro > N / 3sgPro / N-hum / DEM / ProperN / N-anim / N-kin / N-def / N-indef / N_non-hum / N-high_anim / N-low_anim / N-sg / N-pl / 3plPro / N-spec / N-non-specific / N-inanim / N-non-kin / PersN / N-non-sg / Pro-kin / N-non-hum-sg / N-high / N-low / 3duPro / 3sg_humPro / 3sg_non-hum-Pro / Nnon-pers / 3sgPro-high / 3plPro-high / 3sgPro-low / 3plPro-low / NOT-ProperN
P/N-high>N-low	Pro / 3sgPro / N-hum / DEM / ProperN / N-anim / N-kin / N-def / N-high_anim / 2sgPro / 3plPro / 1plPro / N-spec / 1sgPro / 2plPro / PersN / Pro-kin / 1duPro / 2duPro / N-high / 3duPro / 3sg_humPro / 3sg_non-hum-Pro / 3sgPro-high / 3plPro-high > N-indef / N_non-hum / N-low_anim / N-non-specific / N-inanim / N-non-kin / N-non-hum-sg / N-low / Nnon-pers / 3sgPro-low / 3plPro-low / NOT-ProperN
P>N	Pro / 3sgPro / DEM / 2sgPro / 3plPro / 1plPro / 1sgPro / 2plPro / Pro-kin / 1duPro / 2duPro / 3duPro / 3sg_humPro / 3sg_non-hum-Pro / 3sgPro-high / 3plPro-high / 3sgPro-low / 3plPro-low > N / N-hum / ProperN / N-anim / N-kin / N-def / N-indef / N_non-hum / N-high_anim / N-low_anim / N-sg / N-pl / N-spec / N-non-specific / N-inanim / N-non-kin / PersN / N-non-sg / N-non-hum-sg / N-high / N-low / Nnon-pers / NOT-ProperN
nsg>sg	N-pl / 3plPro / 1plPro / 2plPro / N-non-sg / 1duPro / 2duPro / 3duPro / 3plPro-high / 3plPro-low > 3sgPro / 2sgPro / N-sg / 1sgPro / N-non-hum-sg / 3sg_humPro / 3sg_non-hum-Pro / 3sgPro-high / 3sgPro-low
sg>nsg	3sgPro / 2sgPro / N-sg / 1sgPro / N-non-hum-sg / 3sg_humPro / 3sg_non-hum-Pro / 3sgPro-high / 3sgPro-low > N-pl / 3plPro / 1plPro / 2plPro / N-non-sg / 1duPro / 2duPro / 3duPro / 3plPro-high / 3plPro-low

Table 3: A priori defined scales

The principles we test in the following are summarized in Table 3. For example, the 'SAP>3/N' scale predicts that speech act participants rank higher than all other referents, but that languages vary in the mutual ordering of first and second person and that differences in number are irrelevant, while the 'SAP>3>N' in addition predicts differential ranking between pronouns and nouns. The 'P>N' scale reduces this even further. The scale 'P/N-high>N-

low' makes the cut slightly different, capturing mainly effects from animacy, definiteness, specificity and related notions. The table lists two possible ranking of numbers. The sg>nsg ranking is based on the assumption that singular is more indexible than nonsingular and therefore ranks higher: singular items can be better pointed out than multiple items, in the same way as speech act participants can be better pointed at than other referents (Bickel and Nichols, 2007). The reversed ranking nsg>sg is based on the assumption that singular is structurally – and often also morphologically – unmarked relative to nonsingular, and therefore ranks lower (Croft, 1990).

In addition to these theoretically motivated scales, we also explored which scale would fit best with the data empirically. For this, we applied similarity measurements between the referential category types in the universal inventory given in Table 2. For each language, we first examined the referential category types that the alignment set distribution refers to and noted whether a given type is mapped into a structurally marked or a structurally unmarked set (in the sense defined in 2 above), separately for A and O.[4] For example, in order to describe the distribution of alignment sets in Vafsi past tense clauses (cf. the data in 1 and Table 1 above), one needs to refer to its set of pronouns and to a distinction between NPs that rank higher ('N-high') in discourse and NPs that rank lower ('N-low'). Because of the way alignment sets are distributed, the set of pronouns is best liste d explicitly: first person singular, first person plural, second person singular, second person plural, third person singular, and third person plural. With regard to A arguments, the category type 'second person plural' is in the unmarked, all other category types are in the marked set, where they occur in the oblique case; with regard to O arguments, the category types 'second person plural' and 'N-low' are unmarked while all others are marked by the oblique case. Category types like 'N-anim' from Table 2, are not referenced by the alignment split in Vafsi and this is then coded as 'NA' (non-applicable). Keeping unreferenced category types like these would distort the fact that Vafsi makes cuts across types in precisely the way it does. By contrast, if an alignment set has no split at all in a language, we did not code all unreferenced category types as 'NA' but instead we coded them as having the same alignment in the language.

This defines a table where the universal inventory of referential category types from Table 2 is specified as 'marked', 'unmarked' or 'NA'. For each pair of category types, we then computed the relative Hamming distance (also

[4]For how we computed alignment sets, see Section 3.2 below.

known as the Gower coefficient), i.e. the proportion of languages[5] in which the treatment of types as 'marked' vs 'unmarked' differs among the number of languages in which the category types are coded (i.e. are not NA), again separately for A and O. This leads to distance matrices of types for A and O, where each pair within a matrix receives a numerical value between 0 and 1. With 41 category types, the distances between all pairs can be faithfully represented in 40-dimensional space; all lower-dimensional and therefore more interpretable solutions distort the distances t o some extent. In order to find the lowest-dimensional space that still fits the data with an acceptable amount of distortion, we applied non-metrical Multi-Dimensional Scaling (as implemented by Venables and Ripley, 2002). The degree of distortion can be formally measured by what is known the Kruskal Stress value ϕ, which basically expresses the squared deviations of the down-scaled distances from the observed distances, relative to the total of the down-scaled distances. For A arguments, ϕ starts to approach its minimum with 3 dimensions; for O arguments, with 2 dimensions. However, even for one-dimensional solutions, ϕ is modest (ϕ=11% for A; ϕ=16% for O) and detailed inspection of the higher dimensions do not suggest any additional distributional patterns. Specifically and interestingly, no higher-dimensional solution points to, say, a dimension of number as opposed to a dimension of person. Figures 1 and 2 display the one-dimensional solutions.

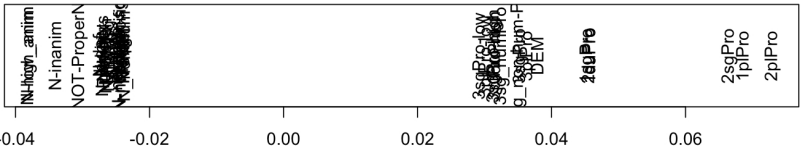


Figure 1: Referential scales as one-dimensional solutions to pairwise comparisons of whether referential categories of A arguments are marked or unmarked across languages. (In order to increase readability, we added a small amount of jittering before plotting; the cluster to the left contains exclusively noun categories.)

[5]When languages split systems between tenses or other non-referential principles, we treat each system as an independent datapoint. See Section 3.2 for discussion.

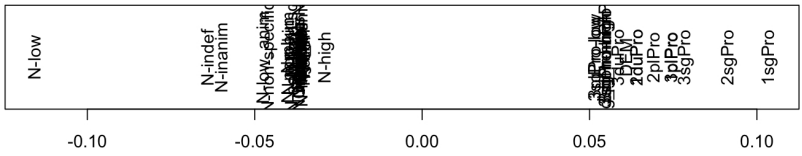


Figure 2: Referential scales as one-dimensional solutions to pairwise comparisons of whether referential categories of O arguments are marked or unmarked across languages. (The cluster to the left contains exclusively noun categories.)

For both A and O arguments, the solutions confirm what is called the P>N scale in Table 3, and this can be taken as an empirical confirmation of its cross-linguistic validity. For A arguments, the solution in addition suggests a scale that ranks first person plural and second person (plural and singular) above other persons:

(4) The 'Non-speaker scale': 1p/2s/2p > 1s/1d/2d/3 > N

This ranking also matches our impressions: first person singular is indeed often treated differently from other persons. We briefly mentioned an example from the Australian language Arrernte above, and Bickel (2000) discusses the special status of first person singular A arguments in a number of Himalayan languages.

For O arguments, the solution in Figure 2 further suggests a scale that ranks first and second person singular, i.e. speaker and addressee, higher than all other persons:

(5) The 'SAPsg scale': 1s/2s > 1d/1p/2d/2p/3 > N

A possible motivation for this is the special exposure of speaker and addressee in communicative events (cf. Bickel et al., 1999; Heath, 1991, 1998). Another possibility suggested by Figure 2 is a special ranking of 'low' on the one hand, and 'indefinite' and 'inanimate' on the other hand; much of this is already captured by the 'high>low' scale defined in Table 3, and Figure 2 does not add much differentiation here. (On the other hand, the results in Figure 2 can of course be seen as an empirical validation of the proposed scale.)

2.4. Two models of scale effects

There are two models of how one can conceive of the way in which scales can determine the distribution of alignment sets. In the model that is traditionally assumed, scales predict a specific distribution of differential argument marking in grammatical systems: each grammatical system with a split either fits or does not fit the prediction, or, formulated as an implicational universal: 'if a language has a split in the case alignment of arguments, this split follows a universal scale'. We call this the 'Type Model'. The alternative, but so far largely unexplored model, is the 'Rank Model': scales are ordered factors of categories that determine the relative probabilities of specific alignment sets for each category. In other words: the odds for case marking on a given argument correlate with the rank of that argument on a universal scale. In the following we discuss how both models can be tested.

Testing Type Models involves counting the number of languages (systems of alignment sets) with alignment splits that fit vs. do not fit the predicted scale. The criterion for fit is made explicit in (6) (assuming the same definition of markedness as in (2) above and the scales as defined in Section 2.3; 'first' means leftmost on the scale and 'position' refers to the set of categories concatenated by '/' in Table 3):[6]

(6) A system of differentially marked alignment sets ℵ fits a scale *X* iff the marked set α of ℵ covers only adjacent positions of *X*, and
   a. for A arguments, α also covers the last element of *X*,
   b. for O arguments, α also covers the first element of *X*.

   A marked set α covers a position X_k iff α occurs in X_k and not-α (the unmarked set(s) of ℵ) does not occur in X_k.

This can be illustrated by the patterns in Table 1 above. For example, given the scale specified in the last column of the table, Spanish fits for O arguments, but there is no prediction for A arguments; Dyirbal fits for both arguments; etc. Khufi or the past tense system of Bartangi, Vafsi or Talysh do not fit with regard to the A argument. Middle Atlas Berber, Gumbaynggir and the Vafsi past tense system do not fit with with regard to the O argument. Obviously, if languages do not reference any of the category types defined by the scale,

[6] Alternatively, one could define fits by the absence of any 'marked>unmarked' (for A) or 'unmarked>marked' (for O) sequence. This was suggested to us by Taras Zakharko, and it is how we in fact compute fits in our statistical report below.

e.g. if a language does not mark number as defined by the sg>nsg scale, the fit cannot be evaluated. In general, a language can be evaluated with regard to a scale X only if each position of X (as defined in Table 3), has a non-empty intersection with the category types referenced by the language.

The null hypothesis for Type Models is defined by the base probabilities of fitting the scale by chance. Given a scale of k positions and a binary contrast between structurally marked vs. unmarked alignment sets, there are 2^k ways in which the alignment sets can fall onto the scale, minus the two cases in which there is no split, i.e. all elements of the scale are marked or all are unmarked. Of all possible matches, k-1 fit the scale in the sense of (6), e.g. on a scale X with k=4, i.e. a>b>c>d, the only fits for A arguments are those where [a>b>c], [a>b], or [a] are marked. Thus, the base chances of finding systems that fit a given scale X with k positions are:

(7) $\pi_0(X) = \frac{k(X)-1}{2^{k(X)}-2}$

Testing individual scales then amounts to binomial tests determining whether the proportion of observed fits exceeds π_0 to such an extent that the excess is unlikely to be due to chance (at, for example, an α-level of .01).

The Rank Model is a standard logistic regression model: a scale is an ordered factor that is expected to affect the chances of finding marked alignment sets. Specifically, the hypotheses to be tested are, for a given scale X:

(8) a. For A: $\log(\frac{\pi(\text{marked})}{\pi(\text{unmarked})}) = \alpha - \beta_1 X + \beta_j Y ... + \beta_k Z$

b. For O: $\log(\frac{\pi(\text{marked})}{\pi(\text{unmarked})}) = \alpha + \beta_1 X + \beta_j Y ... + \beta_k Z$

That is, we hypothesize for A arguments, that the odds for marked alignment sets correlate negatively with X, and for O arguments, that the odds for marked alignment sets correlates positively with X. There may be additional factors $Y...Z$, such as areal diffusion, word order type etc., but the model is statistically supported as along as $Y...Z$ do not interact with X, and coefficient β_1 is larger than 0 to a degree that is unlikely due to chance (at an α-level of, say, .01).

3. Testing for universal effects: methods and data coding

3.1. Methods

As suggested by the preceding, Type Models can be evaluated by binomial tests and Rank Models by logistic regression tests. However, as many typol-

ogists have argued, typological distributions – here, of marked vs. unmarked alignment sets – are not only affected by possible structural or cognitive principles – here, scales –, but also by faithful inheritance within families and areal diffusion resulting from language contact. In other words, the chances of finding a specific alignment set on a specific pronoun in a specific language may just as well be determined by the fact that the language inherited its pronoun system from its ancestor language or that the case distribution assimilated to neighboring languages. Therefore, any typological statistical test needs to control for the confounding factors of family relations and areas.

Areal factors can be built into regression models as factors. For Rank Models, this can be done directly. For Type Models, one can examine the extent to which distributional skewings are replicated across different areas (cf. Dryer, 1989). Family effects can be controlled for in the same way as areal factors if there are only few families with relevant data, and most of them contain many members. In all other contexts, different methods are called for (Bickel, 2007), but for the current datasets, this is not necessary, as we will see.

Note that none of the datasets we use are random samples. Therefore, the principles of random-sampling theory are not applicable, and this makes it impossible to use statistical tests based on this theory. Following the suggestions of Janssen et al. (2006), we therefore employ permutation tests, which test the probability of finding the observed distribution under a random reshuffling of the data. For the Type Model, we employ exhaustive permutations, i.e. exact tests; for the Rank Model we create random samplings of permutations and compute likelihood ratios from these (see Bickel, 2007 for detailed discussion).

3.2. Data coding

Our database contains 333 languages. Most of these were surveyed by us, but about one third of entries was taken from earlier work in the AUTOTYP project on typological databases (Nichols, 1992; Bickel and Nichols, in press-a, in press-b). The database does not track alignment sets per se but instead codes each case in each language for the argument roles it covers, specified for various conditions, including referential category types. For example, the database contains entries like ‘Chantyal: ergative on A in all category types; nominative on S in all category types and on O in ‘N-low’; accusative on O

in 'N-high' and 'pronoun'.[7] Alignment sets are then automatically computed for all intersecting category types. In the Chantyal example, the intersection o f pronouns, 'N-high' and the set of all categories defines the tripartite sets {S},{A},{O}; the intersection of 'N-low' and the set of all categories defines the ergative pattern {S,O}, {A}.

20 languages have splits in case usage across a distinction between participle-based and other tense forms, or between past and nonpast tenses. For the sake of hypothesis testing, we enter these systems as independent datapoints into our computations in the same way as we enter two systems of genealogically related languages as independent datapoints. This raises the number of alignment systems to 353. Whether or not there are dependencies between subsystems within a language or between systems within related languages can then be statistically assessed by looking at family-internal distributions, i.e. by controling for family factors in the sense discussed before.

Area and family factors can be best controlled if areas and families are sampled densely. Therefore, we specifically searched for families with scale-based splits.

4. Results

4.1. Genealogical and geographical distribution

Of the 353 systems in the analysis, 51 have splits on A, i.e. differential A marking of any kind (fitting or not fitting scales), and 99 have splits on O, i.e. differential O marking of any kind; 33 systems have both splits at the same time. The distribution of the splits across families is heavily skewed: the bulk of cases (90% in the case of differential A marking, 79% in the case of differential O marking) are concentrated on only 5 families. Tables 4 and 5 list these; the remaining cases (5 in the case of differential A marking, 21 in the case of differential O marking) are isolated instances in the sense that we do not know – and because of descriptive lacunae, often cannot know – whether other members of the family have splits.

The areal distribution of languages with splits is shown in Maps 1 and 2. In both types of splits, but especially in the case of differential O mark-

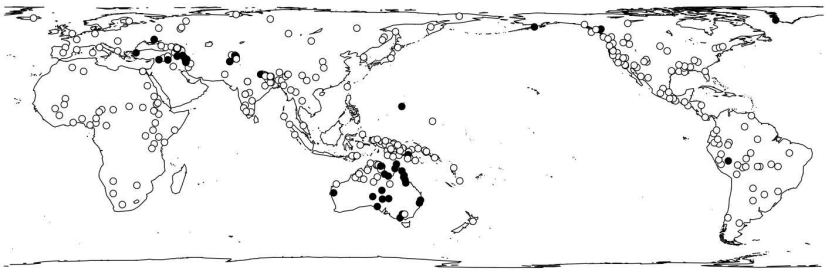
[7] In fact we also code coverage of T and G arguments of ditransitives and lexical splits (e.g. with experiencer verbs). Here, we concentrate on S, A, and O and only survey lexical default classes (open classes).

stock	*N*
Pama-Nyungan	22
Indo-European	16
Nakh-Daghestanian	3
Sino-Tibetan	3
Eskimo-Aleut	2

Table 4: Families with more than one member with differential A marking

stock	*N*
Indo-European	39
Pama-Nyungan	20
Sino-Tibetan	11
Dravidian	6
Turkic	2

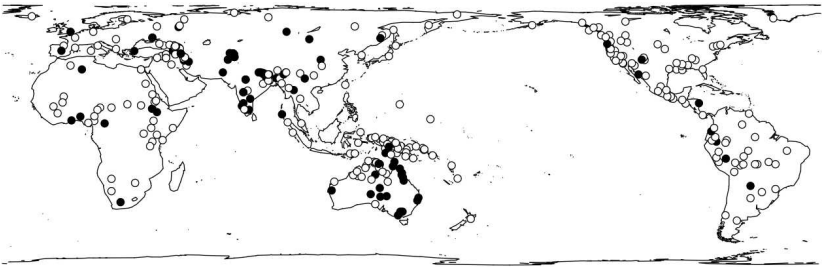
Table 5: Families with more than one member with differential O marking



Map 1: Geographical distribution of languages with differential A marking of any kind (black dots) and languages without differential A marking (white dots)

ing, there are frequency peaks in Eurasia (centered on Indo-Iranian languages, but deeply extending beyond this in the case of differential O marking; Bossong, 1998) and in the New-Guinea-Australia macroareas (centered on Pama-Nyungan languages but extending to Tangkic and Southern New Guinea). For differential O marking, the map suggests the possibility of another macro-areal effect along the northern American Pacific coast and diffusing into South America.

All of these three macro-areas have been noted in previous work (e.g. Bickel and Nichols, 2005*b*, 2006; in press-b; Nichols, 1992, 1993, 1997, 2002). We therefore tested their relative effects on the presence of splits. In order to control for multiplication effects arising from the fact that groups



Map 2: Geographical distribution of languages with differential O marking of any kind (black dots) and languages without differential O marking (white dots)

of languages within areas may be related (e.g. Indo-European within Eurasia) and therefore share a type not because of areal diffusion but because of inheritance, we applied the genealogical sampling algorithm developed in Bickel (2008). This algorithm collapses all those genealogical taxa (at whatever level) that are homogenous to an extent that direct inheritance from their respective proto-language is likely. The collapsed dataset thus represents diachronically independent cases. For example, while overall, Sino-Tibetan and Indo-European show a trend for not having differential A marking, a subst antial number of branches (Kiranti in Sino-Tibetan, Iranian in Indo-European) have differential A marking. Regardless of whether these patterns derive from the proto-languages and all other branches have innovated, or whether the patterns are innovations, while the rest reflects the proto-structure, the distribution must have involved individual diachronic changes. Tables 6 and 7 show that areal effects on these changes are likely since the distributions are statistically significant under a Fisher Exact Test ($p < .001$ for differential A marking, $p = .007$ for differential O marking). Closer inspection of the Pearson residuals, however, suggests that the effects are to a large extent due to Eurasia alone, which has many more cases of differential argument marking than what is expected from the margin totals. Indeed, taken alone, Eurasia accounts for 65% of the χ^2-deviance in differential A marking and for 57% of the χ^2-deviance in differential O marking.[8]

[8]This finding fully converges with the growing body of extra-linguistic evidence suggesting strong and long-standing conditions of population movements that favored the spread of languages, ideas, and objects throughout Eurasia (cf. e.g. Chaubey et al., 2006; Rootsi et al., 2007).

	American	Pacific	Eurasia	NG-Australia	Other	Sum
diff-A		3	21	4	1	29
no diff-A		48	27	53	51	179
Sum		51	48	57	52	208

Table 6: Distribution of differential A of any kind marking across macroareas

	American	Pacific	Eurasia	NG-Australia	Other	Sum
diff-O		6	20	5	8	39
no diff-O		45	45	54	50	194
Sum		51	65	59	58	233

Table 7: Distribution of differential O marking of any kind across macroareas

4.2. The Type Model

For both differential A marking and differential O marking, the number of systems that fit a given scale (in the sense defined in 6 above) outranks the number of systems that do not fit. This is shown in Tables 8 and 9. The total systems evaluated is often smaller than the number of systems with splits because the split may not be relevant for a specific scale (e.g. many languages have differential O marking but for this, number is irrelevant and so there is no split on the number scales); or because a system may not reference the critical category types of a scale (e.g. the language may not at all differentiate number, or a 'high' vs. 'low' distinction). Since simultaneous tests of competing scales on the same dataset increases the risk of familywise error of rejecting true null hypotheses, we applied Holm corrections (Holm, 1979) to the *p*-values obtained from the exact binomial tests.

In order to test whether the distributions hold independently of families, we tested scale fits individually for those families that have many split systems. For both differential A and O marking, this is Pama-Nyungan (Table 10 and 12) and Indo-European (Table 11 and 13). For differential O-marking, Sino-Tibetan is another family contributing a sufficient number of datapoints (more than 10) for statistical testing, but only with regard to the P>N and P/N-high>N-low scales. All other scales are relevant for only 2 out of the 11 languages (both of which are from the Kiranti subgroup: Belhare and Thulung, described in Bickel, 2003 and Lahaussois, 2003, respectively).

The results for differential A marking are as follows: Indo-European (Table 11) alone reveals a significant fit only for the 'Non-Speaker Scale' that was obtained by multi-dimensional scaling in Section 2.3 (Holm-corrected $p = .033$, 11 out of 16 fits) and for the 'nsg>sg' scale ($p = .033$, 13 out of 15

	p-value	fits	percentage	total
1>2>3>N	0.61	9	29%	31
SAP>3>N-high>N-low	0.01	7	70%	10
SAP>3>N	0.01	19	61%	31
SAP>3/N	0.43	20	62%	32
P/N-high>N-low	0.04	14	82%	17
P>N	0.00	43	84%	51
nsg>sg	0.61	18	58%	31
sg>nsg	1.00	3	10%	31
1p/2s/2p>1s/1d/2d/3>N	0.04	18	56%	32

Table 8: Overall fit of systems of differential A marking on various scales; Holm-corrected *p*-values from exact binomial test

	p-value	fits	percentage	total
1>2>3>N	0.00	17	47%	36
SAP>3>N-high>N-low	0.12	8	44%	18
SAP>3>N	0.22	17	47%	36
SAP>3/N	1.00	19	51%	37
P/N-high>N-low	0.00	58	100%	58
P>N	0.00	87	88%	99
nsg>sg	1.00	2	8%	25
sg>nsg	0.22	17	68%	25
1s/2s>1d/1p/2d/2p/3>N	0.00	23	62%	37

Table 9: Overall fit of systems of differential O marking on various scales; Holm-corrected *p*-values from exact binomial test

fits); Pama-Nyungan (Table 10) by contrast only for the P>N ($p < .001$, 21 out of 22 fits) and the P/N-high>N-low scales ($p < .001$, 9 out of 9 fits); there is borderline evidence for the SAP>3>N scale ($p = .062$, 8 out of 11 fits).

The results for differential O marking are as follows: all three families show significant ($p < .05$) evidence for the P>N and P/N-high>N-low scales, with only Indo-European (Table 13) showing a larger number of exceptions (10 exceptions out of 39 splits for the P>N scale). Pama-Nyungan (Table 12) shows additional evidence for the person-differentiating scales 1>2>3>N ($p < .001$, 8 out of 10 fits), SAP>3>N ($p = .02$, 8 out of 10 fits), SAP>3/N ($p = .043$, 9 out of 10 fits) and the scale obtained by Multidimensional Scaling in Section 2.3 (1s/2s>1d/1p/2d/2p/3>N, $p = .02$, 8 out of 10 fits). These scales, in turn, do not yield significant effects in Indo-European, whereas Indo-European shows a trend on the sg>nsg scale ($p = .015$, 16 out of 19 fits). Where the datapoints in Sino-Tibetan are large enough for a statistical assessment, i.e. the P>N and

	p-value	fits	percentage	total
1>2>3>N	0.78	4	36%	11
SAP>3>N-high>N-low	0.20	3	75%	4
SAP>3>N	0.06	8	73%	11
SAP>3/N	0.57	8	73%	11
P/N-high>N-low	0.00	9	100%	9
P>N	0.00	21	95%	22
nsg>sg	1.00	5	45%	11
sg>nsg	1.00	1	9%	11
1p/2s/2p>1s/1d/2d/3>N	0.87	5	45%	11

Table 10: Pama-Nyungan: fit of systems of differential A marking on various scales; Holm-corrected *p*-values from exact binomial test

	p-value	fits	percentage	total
1>2>3>N	1.00	1	7%	15
SAP>3>N-high>N-low	1.00	2	50%	4
SAP>3>N	1.00	7	47%	15
SAP>3/N	1.00	8	50%	16
P/N-high>N-low	1.00	3	60%	5
P>N	0.74	11	69%	16
nsg>sg	0.03	13	87%	15
sg>nsg	1.00	0	0%	15
1p/2s/2p>1s/1d/2d/3>N	0.03	11	69%	16

Table 11: Indo-European: fit of systems of differential A marking on various scales; Holm-corrected *p*-values from exact binomial test

P/N-high>N-low scales, there are significant effects ($p = <.001$, 11 out 11 fits in both cases).

These results suggest that overall, there is statistical evidence for one scale or the other in both differential A and differential O marking. However, this evidence comes only from the distributions within two independent groups of languages: Pama-Nyungan and Indo-European. The statistical signal within Sino-Tibetan constitutes perhaps a third group, but, given the strong areal skewing in differential argument marking that we noted above, we cannot exclude the possibility that the datapoints in Eurasia are areally dependent on each other.

The number-based scales show weak significant effects only in Indo-European (Tables 11 and 13), suggesting a potentially interesting trend for differential A marking to follow a non-singular > singular ranking, which is based on structural markedness, and for differential O marking to follow a

	p-value	fits	percentage	total
1>2>3>N	0.00	8	80%	10
SAP>3>N-high>N-low	0.10	3	75%	4
SAP>3>N	0.02	8	80%	10
SAP>3/N	0.04	9	90%	10
P/N-high>N-low	0.00	8	100%	8
P>N	0.00	19	95%	20
nsg>sg	1.00	2	40%	5
sg>nsg	1.00	0	0%	5
1s/2s>1d/1p/2d/2p/3>N	0.02	8	80%	10

Table 12: Pama-Nyungan: fit of systems of differential O marking on various scales; Holm-corrected *p*-values from exact binomial test

	p-value	fits	percentage	total
1>2>3>N	1.00	7	29%	24
SAP>3>N-high>N-low	1.00	3	25%	12
SAP>3>N	1.00	7	29%	24
SAP>3/N	1.00	8	32%	25
P/N-high>N-low	0.00	25	100%	25
P>N	0.01	29	74%	39
nsg>sg	1.00	0	0%	19
sg>nsg	0.02	16	84%	19
1s/2s>1d/1p/2d/2p/3>N	0.25	13	52%	25

Table 13: Indo-European: fit of systems of differential O marking on various scales; Holm-corrected *p*-values from exact binomial test

singular > nonsingular ranking, which is based on indexibility (cf. Section 2.3). That these trends do not leave stronger signals could be due to their intrinsic weakness, but it could also be due to the fact that number differences do not play a role across all category types, but perhaps only within, say, first and second person arguments. A case in point is the Khufi participle-based tense system, discussed in Section 2.2. above. This system shows a number effect for speech act participants insofar as O arguments are marked in the singular but not in the plural. When testing this as a type on the number scale, Khufi is a counterexample because all thir d persons enter a double-oblique alignment and there are therefore both marked singular and nonsingular O arguments. This violates a scale predicting marked O on singular and unmarked O on nonsingular arguments.

In order to better capture this and similarly-structured distributions, one could add yet further scales to the tests, differentiating between number dis-

tinctions across person categories, but this will ultimately lead to a 'fishing expedition', defeating the very idea of statistical hypothesis testing. Instead of this, we explore Rank Models, which are better equipped to track signals from unevenly distributed oppositions.

4.3. The Rank Model

In a Rank Model, a system like the Khufi participle-based tense is treated as follows: the difference between singular and plural speech act participants enters the analysis by different rank codings: for the singular > nonsingular scale, all singular pronouns are assigned rank 1, all dual and plural pronouns rank 2. The fact that all third person pronouns are (structurally) marked regardless of number is registered by the fact that they are all coded as having marked O arguments. For the regression test, speech act participants will increase the correlation between rank and markedness because only rank 1 is associated with marked O arguments; by contrast, third person O arguments will lower the correlation because they are marked on both rank 1 (singular) and rank 2 (non-singular). In the same way, systems without a split will have the same markedness response on all ranks of the scale. Obviously, if many languages are like this, the odds for marked A and O arguments will be independent of the rank; in other words, the ranks, and with it, the scale on which they are defined, are not a significant predictor of markedness, and markedness is instead perhaps better predictable from other factors. Therefore, Rank Models test for the effect of scales against the base probabilities of any kind of markedness distribution.

Given this design, we enter all systems into the analysis, regardless of whether they have splits or not. For this, we coded each rank of each system for whether or not A and O are marked. If the language does not have a split, all ranks will show the same structural markedness response. The only circumstance where a category is not assigned a rank on a scale, is when it is indeterminate, e.g. a category type like 'animate noun' cannot be assigned a rank on a number-differentiating scale.

In order to control for area factors, we entered a two-way distinction between languages within Eurasia and languages outside Eurasia, in line with the areal findings reported in Section 4.1 above. For controling family effects, we entered all families into the model for which we have at least 5 different systems, regardless of whether there are split or not. These are Indo-European (55 systems), Pama-Nyungan (27), Sino-Tibetan (20), Austronesian (11), Nakh-Daghestanian (9), Dravidian (8), Uto-Aztecan (7), Austroasi-

atic (5), and Uralic (5). Since areas and stocks are naturally correlated, their interactions are not entered into the regression model. Instead, we test the effects of each scale in models together with the area factor, the family factor, and the interactions of each of these with the scale. This leads to a relatively large number of parameters. Combined with a very uneven distribution of factor and interaction levels, there is a relatively high risk of overfittin g and computational problems in maximum likelihood estimation. In response, we used penalized estimation throughout (cf. Harrell, 2001).

Analyzing the odds for structurally marked A arguments revealed no significant effect of the number scales and of the P/N-high>N-low scale, i.e. models with only the area and family factors fit the data just as well as models that also include the scale factor (nsg>sg: $LR_{(1)} = .44$, $p = .539$; sg>nsg: $LR_{(1)} = .43$, $p = .503$; P/N-high>low: $LR_{(1)} = 1.36$, $p = .246$). All other scales showed a significant interaction of the scale factor with the family factor (1>2>3>N: $LR_{(4)} = 17.08$, $p = .031$; SAP>3>N-high>N-low: $LR_{(4)} = 31.14$, $p = .001$; SAP>3>N: $LR_{(3.7)} = 19.76$, $p = .007$; P>N: $LR_{(3)} = 18.06$, $p = .005$; 'non-speaker' scale derived from Multidimensional Scaling: $LR_{(3.7)} = 17.15$, $p = .018$). In the model with the scale SAP>3/N, the interaction between scale and family was only borderline significant ($LR_{(3)} = 9.57$, $p = .079$). However, it is doubtful whether this is sufficient to reject the interaction beyond reasonable doubt because the permutation methods applied here are less powerful than tests based on theoretically assumed distribution. If we make the standard assumption that likelihood ratios follow a χ^2-distribution (Agresti, 2002; Harrell, 2001), the interaction is significant at $p = .024$.

The results for structurally marked O arguments are as follows. For most scales, there is a significant interaction of the scale factor with the family factor (1>2>3>N: $LR_{(5.5)} = 17.24$, $p = .038$; SAP>3>N-high>N-low: $LR_{(5.5)} = 19.12$, $p = .028$; SAP>3/N: $LR_{(3)} = 10.72$, $p = .017$; P>N: $LR_{(3)} = 13.96$, $p < .002$; nsg>sg: $LR_{(3)} = 19.74$, $p < .001$; sg>nsg: $LR_{(2.8)} = 19.04$, $p < .001$). In the case of the SAP>3>N scale, the permutation test revealed no significant interaction ($LR_{(5.4)} = 12.28$, $p = .178$), but this is perhaps due to the weak power of this test: under a χ^2 approximation, the likelihood ratio reached significance ($p = .039$) and it is therefore doubtful whether the interaction should be rejected or not. At any rate, an interaction-free model with the scale fits better than one without the scale ($LR_{(1)} = 29.54$, $p < .001$). Two scal es reach significance without any evidence for interactions with stock or area affiliation: the P/N-high > N-low and the 'SAPsg' scale derived from Multidimensional Scaling. In both cases, models with the scale factor fit the data significantly better than models without (P/N-high>N-low: $LR_{(1)} = 21.43$, $p < .001$; SAPsg: $LR_{(1)} = 39.93$, $p < .001$) and there is no evidence for interactions

(P/N-high>N-low: $LR_{(.6)}$ = .5270, p = .395, $p(\chi^2)$ = .269; SAPsg: LR(5.4) = 10.12, p = .312, $p(\chi^2)$ = .089).

Further analysis suggests that the most parsimonious model for both scales includes the family factor along with the scale factor, but not also the area factor (family factor in the P/N-high>N-low model: $LR_{(7.5)} = 67.88$, $p <$ .001; in the SAPsg model: $LR_{(7.5)} = 111.27$, $p <$ .001; area factor in the P/N-high>N-low model: $LR_{(1)}$ = .01, p = .88; in the SAPgs model: $LR_{(1)}$ = .25, p = .60). The final models are given in (9), with the estimated coefficients of all parameters (X = scale; AN = 'Austronesian', D = 'Dravidian', IE = 'Indo-European', ND = 'Nakh-Dagestanian', PM = 'Pama-Nyungan', U = 'Uralic', UA= 'Uto-Aztecan'; AA 'Austroasiatic' is the (arbitrarily chosen) baseline against which the effects of all other families are compared).

(9) a. For scale *X*: 'P/N-high>N-low':

$$\log\left(\frac{\hat{\pi}(\text{marked O})}{\hat{\pi}(\text{unmarked O})}\right) = -.87 + 1.07X + 1.62AN - 1.90D - .44IE + 3.81ND - .40PM - .91ST - 4.71U - 1.69UA$$

b. For scale *X*: '1s/2s > 1d/1p/2d/2p/3 > N' ('SAPsg scale'):

$$\log\left(\frac{\hat{\pi}(\text{marked O})}{\hat{\pi}(\text{unmarked O})}\right) = -.69 + .77X + 1.57AN - 2.42D - 1.10IE - 4.05ND + .93PM - 1.34ST - 5.23U - 1.70UA$$

This suggests that, independent of families and areas, the odds for O arguments to be marked are $e^{1.07}$=2.91 times higher for pronouns and high-ranking nouns than for low-ranking nouns; and $e^{0.77}$=2.17 times higher on each step of the SAPsg scale.

As shown by Figures 3 and 4, however, the sources of both effects is limited to a subset of the families that were entered into the analysis – other families do not contribute to the effect because the odds for marking O arguments are equal across the scale. In line with this, the overall predictive ability of the models in (9) is fairly limited. Somer's rank correlation between predicted probabilities and observed responses (Harrell, 2001), which ranges between 0 (randomness) and 1 (perfect prediction) is D_{xy} =.486 in the model with the P/N-high>N-low scale (9a) and D_{xy} = .535 in the SAPsg scale (9b). This weak performance is confirmed by comparing the maximum likelihoods of the models against a trivial ('saturated') model, in which each datapoint predicts its own response: the models in (9) are significantly different from the saturated model ($LR_{(231)} = 340.75$, $p <$ .001 for 9a and $LR_{(338)} = 473.2$,

p <.001 for 9b), suggesting that they are unable to predict much of the distribution.[9]

As noted earlier, when compiling the database, we searched specifically for families known to have many split systems. Sampling of further families is therefore more likely to increase the number of families with no splits than the number of families with splits. Adding more families without splits to the database is bound to further decrease the predictive ability of the models in (9): the odds for arguments to be marked are then even less correlated with scale ranks, and are instead likely to be correlated with different factors (family and area affiliation, and perhaps also such factors as word order, since they are known to influence the distribution of case marking.)

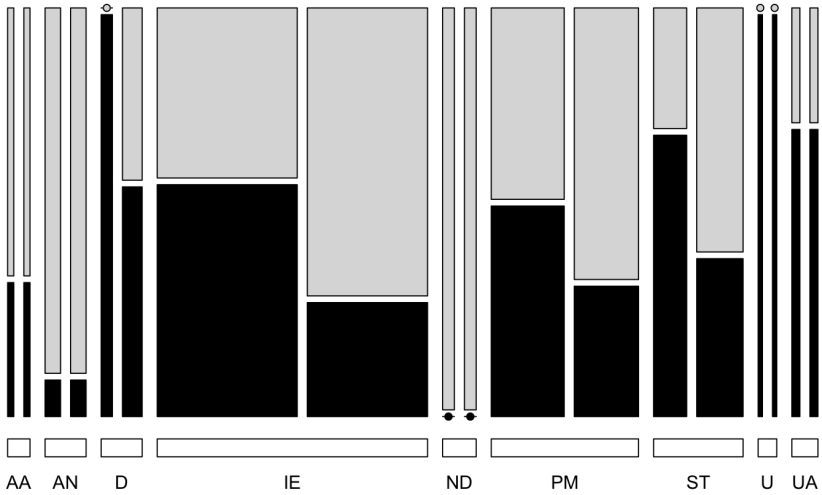


Figure 3: Distribution of marked vs. unmarked O arguments across the P/N-high > N-low scale for all families that entered the analysis (abbreviations as explained in the text). The width of each family-labeling box is proportional to the sample size of the family. Within each family, each bar represents a scale position, arranged from left (highest rank) to right (lowest rank). The width of the bars is proportional to the number of systems under each condition. Within each bar, the black part represents the proportion of marked O arguments (zero is represented by a round circle).

[9]The history of statistical investigations in typology is too young to assess what kinds of model fits can be reasonably expected. In most disciplines, the predictive abilities reported here would probably not qualify as sufficient for accepting a model.

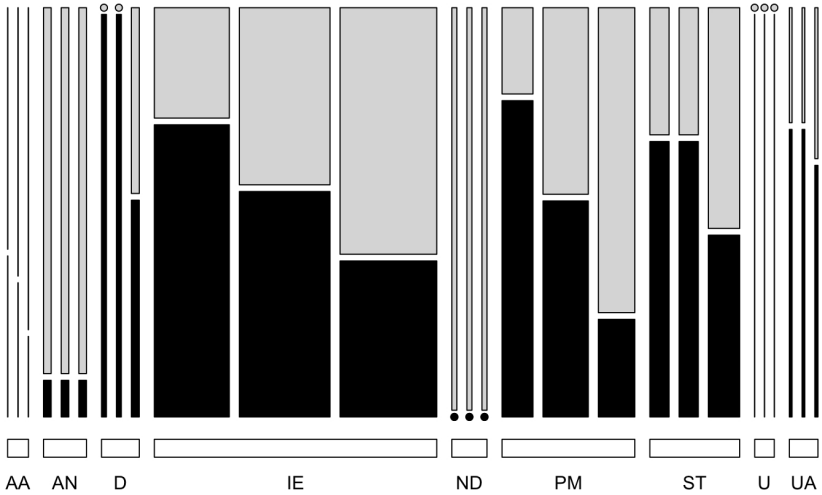


Figure 4: Distribution of marked vs. unmarked O arguments across the SAPsg scale for all families that entered the analysis. Same plotting conventions as in Figure 3

5. Discussion

Examining the Type Model suggests some evidence for scale effects on case marking. However, the evidence is limited to distributional skewings in what amounts to only two or three independent cases: Pama-Nyungan, Indo-European, and Sino-Tibetan. Indo-European and Sino-Tibetan are perhaps only a single independent case because there is a strong Eurasian areality effect on differential argument marking and therefore, we cannot exclude an areal relation between these two families in this regard (also cf. Bossong, 1998 on the Eurasian areality of differential O marking). There are a few further families that show scale effects, but for which we do not have sufficient datapoints in our database (e.g. the Altaic group of families). However, most of these additional families are in Eurasia, and it is therefore unlikely that they are areally independent of Indo-European and Sino-Tibetan. Expanding our database might unearth a few more families with trends towards scale fits, but given that our database is relatively large by current standards, and that we specifically looked out for families with differential argument marking, we doubt that ultimately, more than a handful of cases can be expected. Apart from families with trends towards scale effects, there are also a number of

isolated cases (cf. Section 4.1). To the extent that these have not replicated within families or areas, they point to recessive features, i.e. features that are relatively unstable and prone to loss.

A handful of independent cases of family trends and a set of isolated and possibly recessive cases is an extremely small number for claiming universals, and it is therefore very well possible that the scale effects developed independently of each other, without any common universal principle being involved. Indeed, for most other phenomenon, this would be the default conclusion. For example, a handful families and isolated languages show a trend towards case and number coexponence (Bickel and Nichols, 2005*a*); yet one would not want to derive from this observation universal principles pressuring grammars, along the lines of 'if case is coexponential, then number is a coexponent.' Phenomena like case and number coexponence simply represent types that need to be recognized in inventories of what is possible.

For scale effects, however, it is not even clear whether they are types in this sense because the diversity in how the scales are defined across families and languages suggests that basically each genealogical family (and each isolated case) would constitute its own type, defeating the very purpose of typologizing: for differential A marking, Indo-Iranian shows a skewing towards fitting the 'non-speaker' (1p/2p/2s > other pronouns > nouns)[10] and the nonsingular > singular scale, whereas Pama-Nyungan shows a skewing towards fitting the P>N (pronoun > noun) and the P/N-high > N-low scale — scales that in turn do not yield trend effects in Indo-European. For differential O marking two scales (P>N and P/N-high > N-low) are found relevant for both Pama-Nyungan and Indo-European (and also for Sino-Tibetan), but Pama-Nyungan in add ition shows its own skewings based on person-differentiating scales of which the P>N scale is a special case. For the same person-differentiating scales, however, there is no statistical evidence in Indo-European or Sino-Tibetan O marking, which follow what seems to be a Eurasian standard of definiteness and/or animacy-based splits. Therefore, there is no straightforward way in which the trends across families could be generalized in the format of an implicational universal. Ultimately, the implications will have to be family-specific or area-specific: 'if a language has a split, it follows the scale(s) X_1 in family F_1, but the scale(s) X_2 in family F_2' or 'if a language has a split, it follows the scale(s) X_1 in area A_1, but the scale(s) X_2 in area A_2'. By contrast, what does seem to be a genuine typological type, independent of

[10]This has possibly a more local areal distribution involving the Himalayas and adjacent areas; see Bickel (2000) for tentative suggestions.

families, is the presence or absence of differential argument marking in the abstract; the details are language-specific or family-specific.

This suggests a distributional scenario that does not invoke universal pressure. Under this scenario, differential argument marking developed once in a few languages, in varied shapes. It spread throughout Eurasia, but without the long-standing areal pressure such as the one characterizing this continent, the feature seems to be recessive and apparently did not spread widely, or only locally and in a less consistent fashion (for example in Australia; cf. Tables 6 and 7). The pattern perhaps resulted from uneven case innovation across nouns and pronouns due to earlier phonologically-induced case loss on (some or all) nouns (cf. Filimonova, 2005); or from a side effect of reanalyzing instrumental case as ergatives on inanimate nouns (cf. Garrett, 1990); or from reanalyzing case affixes as markers of definiteness in response to areal diffusion of an abstract definite vs. indefinite opposition — obviously, there are many ways in which differential argument marking may have developed. Once the system was established in a proto-language, it replicated within the family according to whatever further local diachronic changes affected the case systems. This leads to the diversity that we observe.

The scenario just sketched is confirmed by the evidence from the Rank Model: for predicting whether arguments are assigned a marked or unmarked case, scales play no role at all (for some scales), or no role independent of families (for other scales). For differential O marking, two scales do have a significant effect, but the resulting model does not perform well in actually predicting case marking. The reason for these finding is the same as for why the Type Model does not produce evidence for scales: the number of languages where scales leave a trace is very small and much smaller than the number of languages where the distribution of case marking is completely unaffected by scales. And since the effects of many scales interact statistically with what families they operate in, it is likely that the distribution of case marking depends on the specific histories of each family, and not on universal principles.

6. Conclusion

Surveying a dataset of over 350 distinct alignment systems reveals no evidence for a universal trend of scale effects on case marking. There are a few families that do show such effects, but there is no evidence that these effects are based on a shared universal principle. If there were a universal principle, we would expect it (a) to leave signals in many more families, and statistically

independent of these, and (b) to be based on universal principles of scale organization. What we find instead are family-specific scales and a strong areal effect in Eurasia.

Thus, rather than being "one of the most robust generalizations" (Aissen 1999), scale effects on case marking are at best weak generalizations that can be made for a few families. Obviously, this does not entail that one cannot in principle hypothesize higher-level generalizations, but so far we have no empirical evidence for such generalizations; they cannot be taken as a given. We submit that any further exploration of such generalizations requires detailed diachronic studies on how differential argument marking developed in the families that have them. Given the evidence from the areal distributions discussed above, such studies would have to specifically also look into language contact effects leading to the spread of differential argument marking in Eurasia, following the lead of Bossong (1998).

Bibliography

Agresti, Alan (2002): *Categorical data analysis*. Wiley-Interscience, New York.

Aissen, Judith (1999): 'Markedness and subject choice in Optimality Theory', *Natural Language and Linguistic Theory* **17**, 673 – 711.

Bickel, Balthasar (2000): 'Person and evidence in Himalayan languages', *Linguistics of the Tibeto-Burman Area* **23**, 1 – 12.

Bickel, Balthasar (2003): Belhare. *In:* G. Thurgood and R. J. LaPolla, eds, *The Sino-Tibetan languages*. Routledge, London, pp. 546 – 70.

Bickel, Balthasar (2007): A general method for the statistical evaluation of typological distributions. Ms. University of Leipzig, http://www.uni-leipzig.de/∼bickel/papers).

Bickel, Balthasar (2008): 'A refined sampling procedure for genealogical control.', *Sprachtypologie und Universalienforschung* **61**, 221–233.

Bickel, Balthasar (in press-a): Grammatical relations typology. *In:* J. J. Song, ed., *The Oxford Handbook of Language Typology*. Oxford University Press, Oxford.

Bickel, Balthasar (in press-b): On the scope of the referential hierarchy in the typology of grammatical relations.. *In:* G. G. Corbett and M. Noonan, eds, *Case and grammatical relations: papers in honor of Bernard Comrie*. Benjamins, Amsterdam.

Bickel, Balthasar and Johanna Nichols (2002): Autotypologizing databases and their use in fieldwork. *In:* P. Austin, H. Dry and P. Wittenburg, eds,

Proceedings of the International LREC Workshop on Resources and Tools in Field Linguistics, Las Palmas, 26 - 27 May 2002. MPI for Psycholinguistics, Nijmegen.

Bickel, Balthasar and Johanna Nichols (2005*a*): Exponence of selected inflectional formatives. *In:* M. Haspelmath, M. S. Dryer, D. Gil and B. Comrie, eds, *The world atlas of language structures*. Oxford University Press, Oxford, pp. 90 – 93.

Bickel, Balthasar and Johanna Nichols (2005*b*): Inclusive/exclusive as person vs. number categories worldwide. *In:* E. Filimonova, ed., *Clusivity*. Benjamins, Amsterdam, pp. 47 – 70.

Bickel, Balthasar and Johanna Nichols (2006): 'Oceania, the Pacific Rim, and the theory of linguistic areas', *Proceedings of the 32nd Annual Meeting of the Berkeley Linguistics Society* pp. [PDF available at http://www.uni–leipzig.de/∼bickel/research/papers].

Bickel, Balthasar and Johanna Nichols (2007): Inflectional morphology. *In:* T. Shopen, ed., *Language typology and syntactic description*. Cambridge University Press (Revised second edition), Cambridge.

Bickel, Balthasar and Johanna Nichols (in press-a): Case-marking and alignment. *In:* A. Malchukov and A. Spencer, eds, *The Handbook of Case*. Oxford University Press, Oxford.

Bickel, Balthasar and Johanna Nichols (in press-b): The geography of case. *In:* A. Malchukov and A. Spencer, eds, *The Handbook of Case*. Oxford University Press, Oxford.

Bickel, Balthasar and Martin Gaenszle (2007): Generics as First Person Undergoers and the Political History of the Southern Kirant. Paper presented at the 7th Biannual Meeting of the Association for Linguistic Typology, Paris, September 26, 2007.

Bickel, Balthasar, Walter Bisang and Yogendra P. Yādava (1999): 'Face vs. empathy: the social foundations of Maithili verb agreement', *Linguistics* **37**, 481 – 518.

Bossong, Georg (1998): Le marquage différentiel de l'objet dans les langues de l'Europe. *In:* J. Feuillet, ed., *Actance et valence dans les langues de l'Europe*. Mouton de Gruyter, Berlin, pp. 259 – 294.

Chaubey, Gyaneshwer, Mait Metspalu, Toomas Kivisild and Richard Villems (2006): 'Peopling of South Asia: investigating the caste-tribe continuum in India', *BioEssays* **29**, 91 – 100.

Comrie, Bernard (1981): *The languages of the Soviet Union*. Cambridge University Press, Cambridge.

Croft, William (1990): *Typology and universals*. Cambridge University Press, Cambridge.

DeLancey, Scott (1981): 'An interpretation of split ergativity and related patterns', *Language* **57**, 626 – 657.

Dixon, R.M.W. (1972): *The Dyirbal language of North Queensland*. Cambridge University Press, Cambridge.

Dixon, R.M.W. (1994): *Ergativity*. Cambridge University Press, Cambridge.

Dowty, David R. (1991): 'Thematic proto-roles and argument selection', *Language* **67**, 547 – 619.

Dryer, Matthew S. (1989): 'Large linguistic areas and language sampling', *Studies in Language* **13**, 257 – 292.

Dryer, Matthew S. (1992): 'The Greenbergian word order correlations', *Language* **68**, 81 – 138.

Eades, Diana (1979): Gumbainggir. *In:* R. M. W. Dixon and B. J. Blake, eds, *Handbook of Australian Languages 1*. John Benjamins, Amsterdam, pp. 244–361.

Filimonova, Elena (2005): 'The noun phrase hierarchy and relational marking: problems and counterevidence', *Linguistic Typology* **9**, 77 – 113.

Garrett, Andrew (1990): 'The origin of NP split ergativity', *Language* **66**, 261 – 296.

Haig, Geoffrey L. J (2008): *Alignment change in Iranian languages: a construction grammar approach*. Mouton de Gruyter, New York.

Harrell, Frank E. (2001): *Regression modeling strategies*. Springer, New York.

Haspelmath, Martin (2007): 'Pre-established categories don't exist: consequences for language description and typology', *Linguistic Typology* **11**, 119 – 132.

Haspelmath, Martin (2008): Descriptive scales versus comparative scales. *In:* M. Richards and A. L. Malchukov, eds, *Scales*. Vol. 86 of Linguistische Arbeitsberichte, Universität Leipzig.

Haspelmath, Martin (in press): An empirical test of the Agglutination Hypothesis. *In:* S. Scalise, E. Magni and A. Bisetto, eds, *Universals of language today*. Springer, Berlin.

Haspelmath, Martin, Matthew S. Dryer, David Gil and Bernard Comrie, eds (2005): *The world atlas of language structures*. Oxford University Press, Oxford.

Heath, Jeffrey (1991): Pragmatic disguise in pronominal-affix paradigms. *In:* F. Plank, ed., *Paradigms: the economy of inflection*. Mouton de Gruyter, Berlin, pp. 75 – 89.

Heath, Jeffrey (1998): 'Pragmatic skewing in 1 – 2 pronominal Combinations in Native American Languages', *International Journal of American Linguistics* **64**, 83 – 104.

Holm, S. (1979): 'A simple sequentially rejective multiple test procedure', *Scandinavian Journal of Statistics* **6**, 65 – 70.

Janssen, Dirk, Balthasar Bickel and Fernando Zúñiga (2006): 'Randomization tests in language typology', *Linguistic Typology* **10**, 419 – 440.

Keine, Stefan and Gereon Müller (2008): Differential Argument Encoding by Impoverishment. *In:* M. Richards and A. L. Malchukov, eds, *Scales*. Vol. 86 of Linguistische Arbeitsberichte, Universität Leipzig.

Kiparsky, Paul (2004): Universals constrain change; change results in typological generalizations. Ms. Stanford University (http://www.stanford.edu/∼kiparsky/Papers/cornell.pdf, accessed June 23, 2008).

Lahaussois, Aimée (2003): 'Thulung Rai', *Himalayan Linguistics Archive* **1**, 1 – 25.

Meyer, D., A. Zeileis and K. Hornik (2006): 'The Strucplot Framework: Visualizing Multi-way Contingency Tables with vcd.', *Journal of Statistical Software* **17**, 1 – 48.

Moravcsik, Edith (1978): 'On the distribution of ergative and accusative patterns', *Lingua* **45**, 233 – 279.

Morphy, Frances (1983): Djapu, a Yolngu dialect. *In:* R. M. Dixon and B. J. Blake, eds, *Handbook of Australian Languages 3*. John Benjamins, Amsterdam, pp. 1–188.

Nichols, Johanna (1992): *Language diversity in space and time*. The University of Chicago Press, Chicago.

Nichols, Johanna (1993): 'Ergativity and linguistic geography', *Australian Journal of Linguistics* **13**, 39 – 89.

Nichols, Johanna (1997): 'Modeling ancient population structures and population movement in linguistics and archeology', *Annual Review of Anthropology* **26**, 359 – 384.

Nichols, Johanna (2002): The first American languages. *In:* N. G. Jablonski, ed., *The First Americans: The Pleistocene Colonization of the New World*. University of California Press, Berkeley, pp. 273 – 294.

Pencheon, Thomas G. (1973): *Tamazight of the Ayt Ndhir*. Undena, Los Angeles.

R Development Core Team (2008): 'R: A Language and Environment for Statistical Computing'.

Rootsi, S., L.A. Zhivotovsky, M. Baldovic, M. Kayser, I.A. Kutuev, R. Khusainova, M.A. Bermisheva, M. Gubina, S. Fedorova, A.M. IlumÃd'e, E.K. Khusnutdinova, L.P. Osipova, M. Stoneking, V. Ferak, J. Parik, T. Kivisild, P.A. Underhill and R. Villems (2007): 'A counter-clockwise northern route of the Y-chromosome haplogroup N from Southeast Asia towards Europe',

European Journal of Human Genetics **15**, 204 – 211.
Schulze, Wolfgang (2000): *Northern Talysh*. Lincom Europa, Munich.
Silverstein, Michael (1976): Hierarchy of features and ergativity. *In:* R. Dixon, ed., *Grammatical Categories in Australian Languages*. Humanities Press, New Jersey, pp. 112 – 171.
Sokolova, Valentina Stepanovna (1959): *Rusanskie i chufskie teksty i slovar'*. Nauka, Moskva.
Stilo, Donald (2004): *Vafsi Folk Tales*. Dr. Ludwig Reichert Verlag, Wiesbaden.
Venables, W. N. and Brian D. Ripley (2002): *Modern applied statistics with S*. Springer, New York.
Wilkins, David (1989): Mparntwe Arrernte (Aranda): studies in the structure and semantics of grammar. PhD thesis, Australian National University.